Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 8843113, 13 pages
https://doi.org/10.1155/2020/8843113

Research Article

Person Detection for an Orthogonally Placed Monocular Camera

Pavel Skrabanek,[1] **Petr Dolezel**,[2] **Zdenek Nemec**,[2] **and Dominik Stursa**[2]

[1]*Faculty of Mechanical Engineering, Brno University of Technology, Brno, Czech Republic*
[2]*Faculty of Electrical Engineering and Informatics, University of Pardubice, Pardubice, Czech Republic*

Correspondence should be addressed to Pavel Skrabanek; pavel.skrabanek@vut.cz

Received 6 May 2020; Revised 15 July 2020; Accepted 23 September 2020; Published 14 October 2020

Academic Editor: Kun Xie

Counting of passengers entering and exiting means of transport is one of the basic functionalities of passenger flow monitoring systems. Exact numbers of passengers are important in areas such as public transport surveillance, passenger flow prediction, transport planning, and transport vehicle load monitoring. To allow mass utilization of passenger flow monitoring systems, their cost must be low. As the overall price is mainly given by prices of the used sensor and processing unit, we propose the utilization of a visible spectrum camera and data processing algorithms of low time complexity to ensure a low price of the final product. To guarantee the anonymity of passengers, we suggest orthogonal scanning of a scene. As the precision of the counting is relevantly influenced by the precision of passenger recognition, we focus on the development of an appropriate recognition method. We present two opposite approaches which can be used for the passenger recognition in means of transport with and without entrance steps, or with split level flooring. The first approach is the utilization of an appropriate convolutional neural network (ConvNet), which is currently the prevailing approach in computer vision. The second approach is the utilization of histograms of oriented gradients (HOG) features in combination with a support vector machine classifier. This approach is a representative of classical methods. We study both approaches in terms of practical applications, where real-time processing of data is one of the basic assumptions. Specifically, we examine classification performance and time complexity of the approaches for various topologies and settings, respectively. For this purpose, we form and make publicly available a large-scale, class-balanced dataset of labelled RGB images. We demonstrate that, compared to ConvNets, the HOG-based passenger recognition is more suitable for practical applications. For an appropriate setting, it defeats the ConvNets in terms of time complexity while keeping excellent classification performance. To allow verification of theoretical findings, we construct an engineering prototype of the system.

1. Introduction

In passenger transport, person flow monitoring has an indispensable importance. In some areas of public transport, passenger flow monitoring systems are employed to automate this task. One of the basic measures, which must be provided by the system, is the number of transported passengers. A precise counting of passengers entering and exiting means of transport has a positive effect on public transport surveillance, passenger flow prediction, transport planning, transport vehicle load monitoring, station control and management, and cost optimization [1, 2].

To ensure a robust and precise counting of passengers in real time, a passenger flow monitoring system must be based on an appropriate imaging system and data processing algorithms. In order to allow a mass deployment of such a monitoring system, a low-cost final solution is equally important. The solution should also meet legal requirements where passenger anonymity is of great importance. Specifically, identification of individuals according to their faces must be avoided.

The imaging system must ensure the acquisition and processing of data, i.e., its basic components are a sensor and a processing unit. In order to develop an inexpensive solution, low price of both components is crucial. While the lower price limit of the processing unit is mainly given by the complexities of used data processing algorithms, the lower price limit of the sensor is given by the used sensing

technology. Radar sensors [3], laser scanners [4], 3D laser scanners [5], or infrared sensors [6] are applicable for the counting of passengers. All these sensors naturally guarantee a high level of passenger anonymity. Their main drawbacks are high prices of the sensors and frequent errors in the counting [7, 8]. For these reasons, cameras operating in the visible spectrum of light are preferably used for the counting of persons [9]. Conventional cameras (cameras operating at wavelengths of visible light) are significantly cheaper, compared to the previously mentioned sensors. The cameras can be combined with depth sensing devices [10]. The fusion of data can result in a more balanced trade-off between false positives and false negatives [11]. On the other hand, the depth sensing devices increase the final prices of sensors, i.e., utilization of a depth sensing device would increase the final price of the imaging system.

The automated counting of persons in a scene is usually carried out in colour images or in sequences of colour images. Many data processing algorithms aimed at precise counting of persons in crowded scene images have been presented [9]. Most of them are designed for overriding installations of cameras. Cameras installed at public as well as at private places usually look down on scenes from angles that typically range between 40° and 80° (from the ground). Considering low subject distances in transportation means (a distance between a camera and a passenger), we conclude that the anonymity of passengers is not guaranteed for such a setup (i.e., data processing algorithms aimed at processing of such images cannot be used for the counting of passengers). Only orthogonally captured images (camera placed above a scene, looking directly down on the scene) assure a high level of passenger anonymity (Figure 1).

A data processing chain, aimed at counting of persons in orthogonally captured images, is compounded of three fundamental steps: person detection, multiperson tracking, and person counting (Figure 2). In the first step, a processed image is examined for the presence of persons. The following step is the tracking, where all persons detected within the first step are matched with existing tracking models of persons. In the case a person cannot be associated with any existing model, a new tracking model is initialized. The last step of the chain is the counting. If a person described by a tracking model leaves the scene, which is usually defined by virtual lines, counting is triggered [11]. Naturally, an integral part of this data processing chain is an algorithm which splits video data provided by a camera into individual images.

Accuracy and time complexity of the data processing chain is primarily given by accuracy and time complexity of the person detection. Person detection is a process of location and recognition of persons in images. Within this process, possible locations of persons (regions) are proposed using an appropriate technique. The regions determine candidate object images, which are classified using an appropriate object recognition system. The proposition of regions can be carried out using an exhaustive method such as a sliding window [12] or using an advanced time-efficient method such as a selective search algorithm [13]. In modern object detection systems, both the location and the recognition are carried out by a single deep neural network [14–16]. These systems are characterized by high detection accuracy but high time complexity.

As the analysis shows, a low-cost passenger counting system should be based on a conventional camera (due to low prices of visible light cameras). In order to guarantee the anonymity of passengers, the camera must be placed above a scene, looking directly down on the scene. For the data processing, methods capable of processing orthogonally captured images must be used. The resulting data processing chain must be robust and precise. To keep the low-cost requirement, the time complexity of the methods should be as low as possible. From this perspective, the detection of passengers seems to be the weak link in the chain.

As the time complexity of the single deep neural network detectors is high [14–16], we tend to implement a passenger detector as a two-stage system. When using a robust and time-efficient region proposal method such as selective search algorithm [13], the accuracy and computational complexity of the detector is mainly given by a used object recognition method. In colour images, the recognition of persons typically relies on optical flow features [11, 17, 18]. An alternative approach to the detection of persons is the detection of their heads and shoulders [19]; however, a head itself can provide a strong feature due to its almost circular shape. The counting of heads is typically used by counting of persons in dense crowd images [20, 21]. Recognition of heads in orthogonally captured images can also rely on the optic flow analysis [22]. The main disadvantages of optic flow-based methods are their high computational complexity and noise sensitivity [23].

Considering the importance of passenger recognition for their counting, we focus on the development of a price-competitive and time-efficient object recognition system. As the system is aimed at recognition of passengers, we name it "the passenger recognition system." As the trend in object recognition is still clearly heading towards convolutional neural networks (ConvNets) [24, 25], we examine the performance of ConvNets for passenger recognition. Usually, ConvNet-based object recognition systems have good classification performance, but their time complexity is typically high. For this reason, we propose a competitive approach which is based on histograms of oriented gradients (HOG) features [26] and on a support vector machine (SVM) classifier. For an appropriate setting of parameters, HOG-based object recognition can have good classification performance while keeping low time complexity [27].

Recognition of passengers in orthogonally captured images using the HOG features and the SVM classifier, based on object images which comprise of heads and shoulders of passengers, has proven useful in scenes without height differences [19]. Modern public means of transport are increasingly low-floor (i.e., there is no or negligible height difference in the area of a doorway), but a substantial part of operated buses, trams, trains, and trolleybuses are high-floor [28–30]. Considering this fact, we conclude that the robustness of the HOG-based passenger recognition system must be verified in the context of variable distances between a camera lens and passenger heads. We also consider that the

Figure 1: Examples of images orthogonally captured in a tram. Identification of persons in the images according to their faces is implausible.

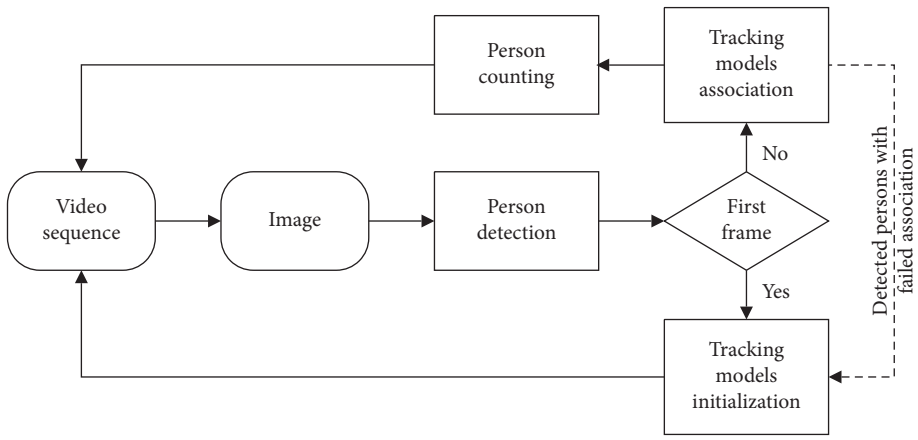


Figure 2: Overall diagram of a person counting system. The system processes images streamed by a monocular camera. Each image is firstly examined for a presence of persons. Positive samples are then associated with existing tracking models of persons. If sample is not associated with any existing model, a new model is initialized. If a tracking model leaves scene, which is defined by virtual lines, counting is triggered.

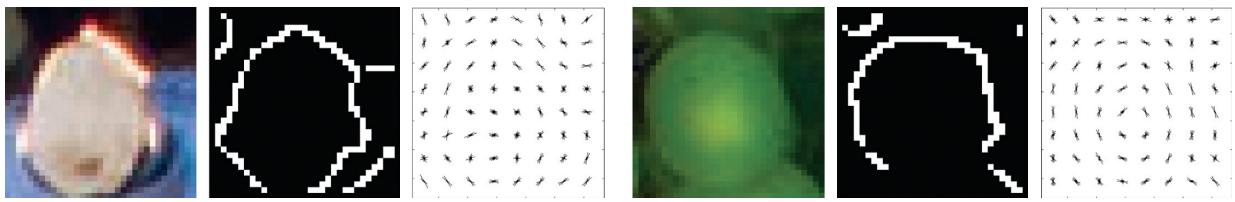
time complexity of the system can be reduced once the object image contains only the heads of the passengers (omitting the shoulders will result in smaller object images and consequently reduce data processing time). We deduce the suitability of such an approach from remarkable results of HOG-based object recognition systems on similar tasks, e.g., for grape detection [31, 32] (see Figure 3; the round shape of grapes is similar to the shape of heads).

Within this article, we study the classification performances and time complexities of passenger recognition systems. The systems are aimed at recognition of passengers in orthogonally captured images, where the recognition quality is not adversely affected by the variable distance between the passenger and camera sensor. The passenger recognition systems are based either on ConvNets or on HOG features. Both approaches rely on the detection of heads. In the case of ConvNet-based systems, we consider ConvNets of various topologies. In the case of the HOG-based system, we examine various settings of parameters. We validate the theoretical results in a real-world application. For this purpose, we develop an engineering prototype of the system.

2. Material and Methods

2.1. Engineering Prototype of the System. Two basic components of the system are the sensor and the processing unit (Figure 4). In our case, we use an industrial colour camera Basler acA2500-60uc [34] as the sensor. The camera is placed in a means of transport, at the ceiling near a door. The optical axis of the camera is perpendicular to the vehicle floor. Considering the construction of means of transport, we expect the average subject distances to be from 0.2 m to 1 m. The camera should monitor an area of about 2.4 m × 2.0 m. With respect to these parameters, we equipped the camera with a Computar M3514-MP lens [35]. The output of the camera (i.e., the input of the data processing chain) is a sequence of RGB images.

We use the prototype for a data collection as well as for the validation of the proposed recognition methods, i.e., the prototype must be capable of processing acquired images in real time. In order to allow testing of all proposed solutions (including solutions based on ConvNets), we use a single-board computer VOB-P3310. It offers an NVIDIA Tegra X2 (2.0 GHz, 6 cores) CPU together with 8 GB RAM and it



(a) (b)

Figure 3: Comparison of head (three images in (a)) and grape images (three images in (b)). For each category, we provide an original RGB image, an image obtained by filtering of the RGB image using the Canny edge detector [33], and gradients obtained using a HOG descriptor [26], respectively.

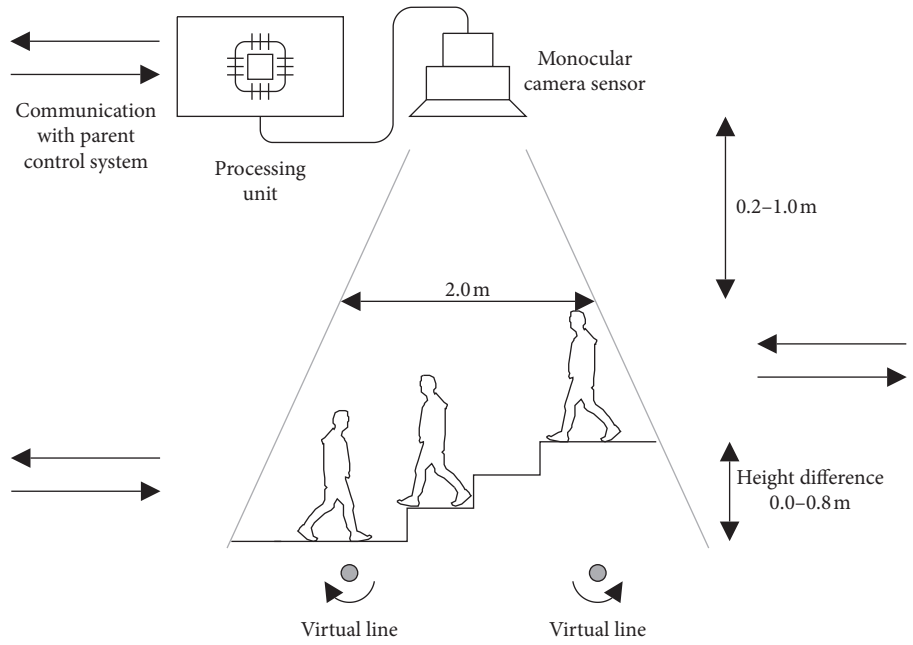


Figure 4: Architecture of person flow monitoring system (side view). Crossing space between lines causes person to be counted.

provides wide communication possibilities (USB 2.0, 3.0, SATA, WiFi) [36].

2.2. Passenger Recognition. Candidate object images may or may not contain complete heads of passengers (Figure 5). According to this criterion, the images are classified either as "head" or "not head" by a passenger recognition system. Inputs of the recognition system are sized normalized RGB object images of dimensions 51×51 pixels ([51, 51] px). Its outputs are labels of the images, where labels "head" and "not head" are allowed.

2.2.1. Passenger Recognition Based on ConvNets. In terms of classification accuracy, the state-of-the-art object recognition systems are based on one of the successful deep ConvNet architectures [37]. Mostly, they process raw image data (i.e., no image preprocessing is usually carried out). They consist of multiple layers arranged in a feed-forward manner. Upper and lower level layers ensure feature extraction and classification of object images, respectively. The feature extraction is usually carried out using convolutional and pooling layers, where the convolutional layers are typically combined with a ReLU activation function. The classification is generally ensured by a softmax activation function. The function processes data at the output of the last network layer, where a fully connected layer is placed. The number of neurons of this layer corresponds to the number of object classes [38]. The main drawback of the state-of-the-art deep ConvNet architectures is their high computational demands.

The passenger recognition can be simply implemented as a ConvNet of an appropriate architecture, where the network ensures both feature extraction and classification (Figure 6). As a low time complexity of the system is crucial, we test the performance of five ConvNet architectures of different complexities.

The simplest architecture, Net1, consists of one convolutional layer (32 filters with 3×3 px kernels), one max-pooling layer (2×2 px nonoverlapping pools), and two fully connected layers of 512 and 2 neurons, respectively. The classification is carried out using the softmax function. In the

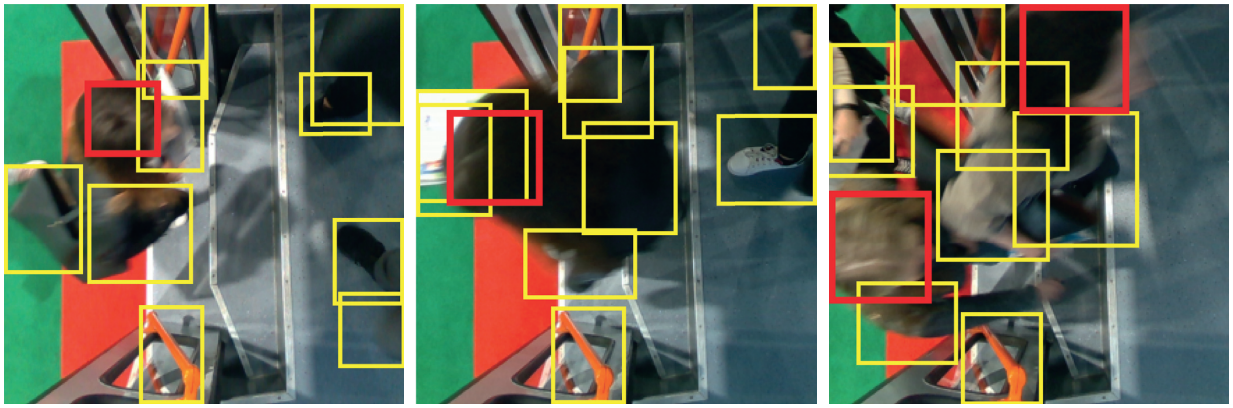


Figure 5: Detection of passengers in orthogonally acquired video data. Candidate object images proposed by a search algorithm (yellow rectangles) are tested. Some of them are classified as "head" (red rectangles), which represents the locations of passengers.

second simplest architecture, Net2, we replace the convolutional and the max-pooling layers by the sequence of layers: convolutional layer (32, 3×3) + convolutional layer (32, 3×3) + max-pooling layer + convolutional layer (64, 3×3) + convolutional layer (64, 3×3) + max-pooling layer, where 2×2 px nonoverlapping pools are used at both pooling layers. In both networks, we use ReLU activation functions at the convolution and fully connected layers. To reduce overfitting, we place dropout layers after each max-pooling layer and after the first fully connected layers in both networks. The dropout rate is 25% and 50% for the max-pooling and the fully connected layers, respectively.

The remaining three architectures studied within this article are the well-known LeNet-5 [39, 40], AlexNet [41], and VGG-16 net [42]. The networks are ordered according to their complexities. The LeNet-5 is the pioneering ConvNet of a relatively simple architecture. AlexNet is probably the most cited deep ConvNet with a huge number of industrial and engineering applications. VGG-16 is a representative of very deep ConvNets. As it consists of only 13 and 3 convolutional and fully connected layers, respectively, the real-time processing of data by VGG-16 implemented in the engineering prototype (Section 2.1) is still possible.

We train all the networks from scratch with initial weights set randomly with normal distribution (mean = 0, standard deviation = 0.05). In addition, we use transfer learning (TL) for AlexNet and VGG-16 in order to test the possibility of better performance [41, 42]. For both architectures, we fine-tune the last three layers of the pretrained networks.

Due to a stochastic character of the training process, we repeat the training a hundred times for each network and training strategy. For each training, we randomly split up a training set into training and validation subsets at the ratio 85 : 15. For each training subset, we run the training in a batch mode for 100 epochs with batches of 32 images. We randomly shuffle data in training subsets for each epoch. We use an ADAM optimizer [43] with initial learning rate setup at 10^{-3} and exponential decay rates for the first and second moment estimates setup at 0.9 and 0.999, respectively. The optimizer and setting of the hyperparameters are the results of a pilot study. We minimize a binary cross-entropy function:

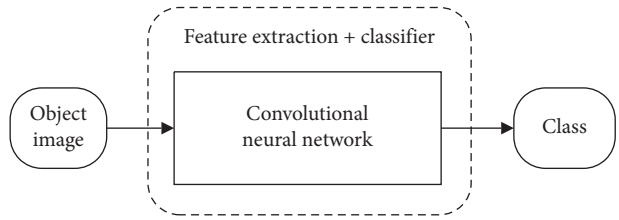


Figure 6: Vision pipeline of the passenger recognition systems based on ConvNets.

$$E_{\text{ConvNet}} = -\frac{1}{n}\sum_{j=1}^{n}\left[\hat{y}_j \ln\left(y_j\right) + \left(1 - y_j\right)\ln\left(1 - \hat{y}_j\right)\right], \quad (1)$$

where n is the number of images in the training subset and y_j and $\hat{y}_j$ are an actual and a predicted class of the j-th object image, respectively. We validate each such trained network on the corresponding validation subset using the cross-entropy function (1).

2.2.2. Passenger Recognition Based on HOG and SVM. Herein, we present a passenger recognition system developed using traditional computer vision techniques. A vision pipeline of the system consists of three successive steps: image preprocessing, feature extraction, and classification (Figure 7). For the feature extraction and classification, we use the HOG descriptor and SVM classifier, respectively. In order to reduce the time complexity of the system, we convert input RGB images to the grayscale format within the image preprocessing. The conversion is carried out according to the ITU-R recommendation BT.601 [44]. The second step of the preprocessing is the unity-based normalization of the grayscale images [31].

The HOG descriptor encodes local shape information from regions within an image into a feature vector [26]. The descriptor has five parameters: number of bins, orientation binning, size of cells (in pixels), number of cells in blocks, and number of overlapping cells between adjacent blocks. As the size of cells significantly influences the final performance of image recognition systems [27] (Figure 8), we study the

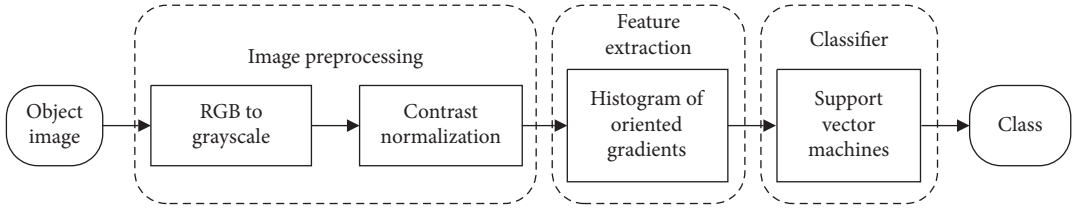


Figure 7: Vision pipeline of the passenger recognition system based on the HOG descriptor and on the SVM classifier. The conversion combined with the normalization is aimed at reducing of time complexity of the system.

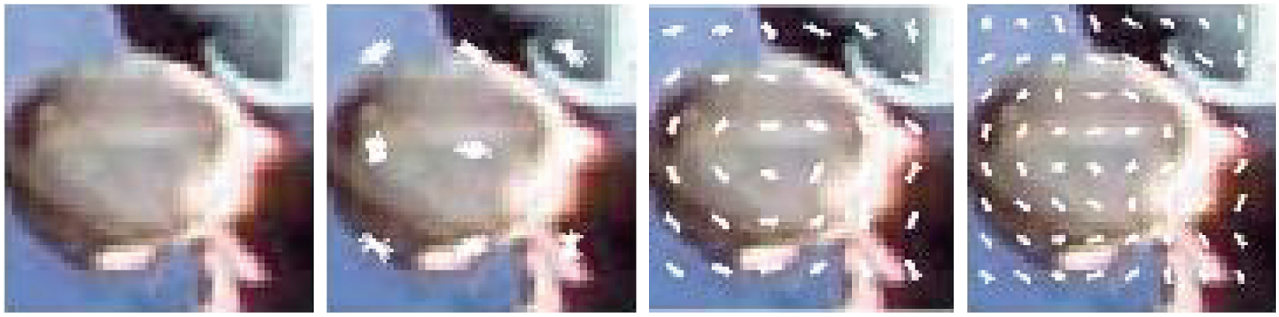


Figure 8: From left to right: the original object image [51, 51] px and images with highlighted gradients of HOG features for cell of sizes [16, 16] px, [8, 8] px, and [6, 6] px, respectively. The length of white abscissae is related to the gradients in the image.

influence of this parameter on the classification performance of the HOG-based passenger recognition system. Specifically, we consider cells of sizes $[6, 6], [8, 8], \ldots, [16, 16]$ px. For the remaining parameters, we use a conservative setting which has proven to be efficient: linear gradient voting into 9 bins linearly spread over 0 to 180 degrees, blocks of 2×2 cells, and 1 overlapping cell between adjacent blocks in both directions.

Training of the SVM classifier is an optimization problem which searches for a hyperplane with a maximal margin from the training data [45]. In the case that the data is not linearly separable, the data must be transformed into a linearly separable problem using an appropriate kernel function. For strongly nonlinear problems, selection of the kernel function is crucial. Considering this fact, we test the influence of various kernels on the performance of the HOG-based passenger recognition system. Specifically, we focus on the well-established linear, Gaussian radial basis function (RBF), and polynomial kernel functions (we use polynomial kernel with order equal to 2 and 3).

Performances of SVM classifiers are also influenced by settings of their regularization constants. In the case that an SVM classifier uses the RBF kernel, its performance is further influenced by kernel width. In a pilot study, we have found setting of the regularization constant at 1 to be optimal. We use a subsampling-based heuristic procedure to find the optimal setting of the kernel width.

As classification performances of classifiers strongly depend on the composition of training sets, we search for a setting ensuring the best performance of the HOG-based passenger recognition system. We carry out the search in the manner described in Section 2.2.1. Specifically, we randomly split up the training set into training and validation subsets at the ratio 85 : 15, and we train and validate the system on the subsets. We repeat the training-validation process a hundred times for each possible combination of kernel function and cell size. We carry out the validation on corresponding validation subsets using a loss function that is given as a sum of misclassified observations, i.e.,

$$E_{\text{SVM}} = \sum_{j=1}^{n} I\left\{\hat{y}_j \neq y_j\right\}, \tag{2}$$

where $I\{\cdot\}$ is the indicator function.

2.3. Evaluation of Passenger Recognition. Two key aspects of the presented passenger recognition systems are their classification performances and their time complexities. A common practice of the evaluation of the classification performance is calculation of accuracy over a testing set (a dataset independent of the training set). For the classification of images into categories "positive" and "negative," the accuracy is given as follows:

$$\text{accuracy} = \frac{|\text{TP}| + |\text{TN}|}{|\text{TP}| + |\text{FP}| + |\text{TN}| + |\text{FN}|}, \tag{3}$$

where $|\text{TP}|$ is the number of correctly classified positive images, $|\text{FN}|$ is the number of misclassified positive images, $|\text{FP}|$ is the number of misclassified negative images, and $|\text{TN}|$ is the number of correctly classified negative images.

To evaluate the classification performance comprehensively, we use three additional measures [31, 46]:

$$\text{precision} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FP}|}, \tag{4}$$

$$\text{recall} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|}, \tag{5}$$

$$F1 - \text{score} = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}}. \quad (6)$$

To evaluate the time complexities of the systems, we measure times that the systems needed to process the testing set. To keep the results independent on the used hardware, we operate with a relative computational time. For the j-th evaluated system, its relative computational time is given as follows:

$$T_j = \frac{t_j}{\max\{t_1, t_2, \ldots, t_k\}} \times 100, \quad (7)$$

where t_j is time the j-th system needs to process the data and k is the number of all evaluated systems.

We carry out the evaluation of passenger recognition systems using the best models obtained within the training process (see Sections 2.2.1 and 2.2.2). In the case of ConvNet-based systems, we use for each architecture, the model with the smallest value of the cost function (1) obtained by its validation. In the case of the HOG-based system, we use for each setting the model with the smallest value of the cost function (2) obtained by its validation.

2.4. Training and Testing Sets. Quality and composition of the training and testing sets conspicuously influence the overall performance of object recognition systems in real-life applications. Data included in the sets should reflect as many aspects of the real situation as possible. Considering this fact, we base the sets on video sequences acquired in the means of public transport and similar public places under various light conditions, using the engineering prototype.

A set of candidate object images generated by a search algorithm from a frame is imbalanced (often highly) [12, 13] with a predominance of images without complete heads (Figure 5). As conventional SVMs are not suitable for the imbalanced learning tasks [47], the training and testing sets must be balanced to get unbiased results. Considering these facts, we create the sets manually to ensure the balance of the classes in the sets.

Specifically, we perform four distinct video recording experiments. They are set to simulate the real situation as well as to comprehend the architecture of the assumed person flow monitoring system (see Figure 4). All the experiments include stairs and a group of persons walking under the acquisition sensor. Men, women, and children as well as people with and without a head cover (hats, scarves, caps, and hoods) are included. Since the used camera lens is focused manually (once for each experiment), the acquired frames show certain blurring according to the specific distance between the object and the lens. We varied locations, lighting conditions, number of frames, mean distance between persons $\overline{D}_{\text{PP}}$ (mean distance between a subject and two other nearest persons), and minimal and maximal distances between a head and the sensor, $\min D_{\text{HS}}$ and $\max D_{\text{HS}}$, in each experiment (Table 1).

We cut out and size normalize 6020 unique object images from the video data (dimension of the normalized images are $[51, 51]$ px). We label the images according to the presence/absence of heads (Figure 9). We mix and divide the labelled images into the training and testing sets according to Table 2. We make the sets publicly available at [48]. The sets contain large-scale class-balanced data which make them universally applicable (the sets can be used to design any classifier including classifiers, which are not suitable to be trained with imbalanced training sets).

3. Results

3.1. Validation of Passenger Recognition Systems. We train and validate each proposed architecture (ConvNet-based system) and each setting (HOG-based system) a hundred times. To show the validation results, we use box plots. Results obtained for the systems based on ConvNets are shown in Figure 10. The central lines in the graphs are medians of the loss function (1); the edges of the boxes are 25th and 75th percentiles; and the whiskers indicate the variability outside the upper and lower quartiles. The data are grouped according to the architectures and training strategies (x-axis). The values on the y-axis correspond to the loss function values.

Figure 11 shows validation results obtained for the HOG-based passenger recognition system using the loss function (2). We use a separate graph for each kernel function. Data in the graphs are grouped according to the sizes of cells. Outliers are symbolized using stars.

3.2. Classification Performance of Passenger Recognition Systems. In Table 3, we summarize evaluation results obtained from the testing set using the measures (3)–(6). The results are grouped into two sections according to the approach they are based on. The best results obtained for each measure are in bold for both approaches.

3.3. Time Complexities of Passenger Recognition Systems. We display relative computational times (7) as a bar graph (the lower chart in Figure 12), where the time and evaluated systems are on the y- and x-axes, respectively. Above each result, we display the $F1$-score (6) of the system as a bar graph (the upper chart in Figure 12), where the $F1$-score and evaluated systems are on the y- and x-axes, respectively.

4. Discussion

The main objective of the presented work is comparison of the two well-established object recognition approaches for the passenger recognition task. As the evaluation results (Table 3) show, for the cells of size $[10, 10]$ px and the polynomial kernel function of degree 3, the classification performance of the HOG-based system slightly exceeds the classification performance of ConvNet-based systems. For this setting, the HOG-based system has the highest values of all four measures. The ConvNet-based systems show the best results for only one measure at a time (aside from LeNet-5 with highest accuracy and $F1$-score). Except for recall, the HOG-based system also exceeds the ConvNet-based systems in sizes of the performance measure values. Further, for this

Table 1: Parameters of the video recording experiments. For each experiment, a location, lighting conditions, number of frames, mean distance between persons $\overline{D}_{PP}$, and minimal and maximal distances between a head and the sensor, min D_{HS} and max D_{HS}, are specified. Note that individual persons can be present in multiple frames in different positions.

Location	Light	No. of frames	$\overline{D}_{PP}$	min D_{HS}	max D_{HS}
Outdoors	Ambient, strong	1720	0.00–0.50	1.2	1.6
Indoors	Ambient, strong	1700	0.25–0.75	0.6	1.8
Indoors	Ambient, weak	1400	0.25–0.75	0.4	1.4
Indoors	Artificial, weak	1200	0.50–1.00	0.2	1.0

(a)

(b)

Figure 9: Examples of object images in the sets. The first three images (a) are labelled as "head" while the remaining three (b) are labelled as "not head."

Table 2: Dataset.

Set	Training		Testing	
Class	"Head"	"Not head"	"Head"	"Not head"
No. of images	2008	2012	1000	1000

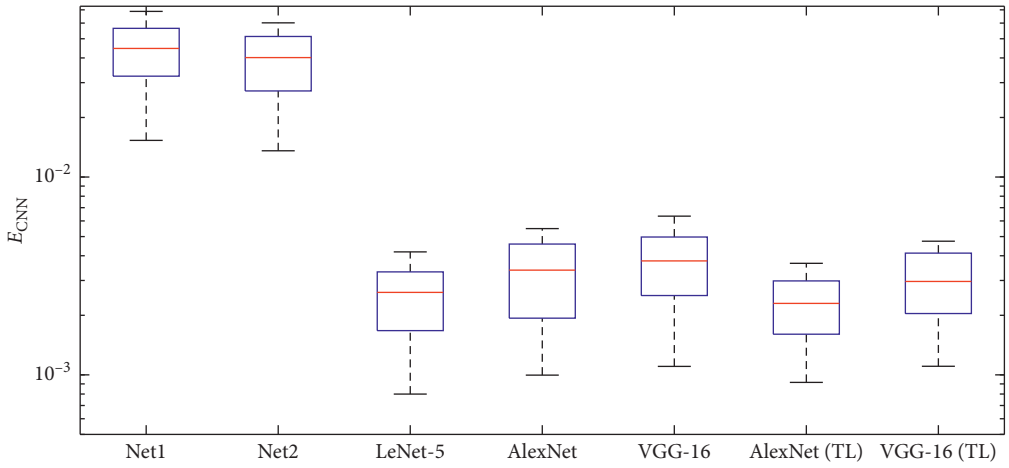


Figure 10: Boxplot representation of loss function values (y-axis) obtained by validation of ConvNet-based passenger recognition systems. Architectures of the ConvNets and training strategy are at x-axis (TL = transfer learning, otherwise, the network was trained from scratch).

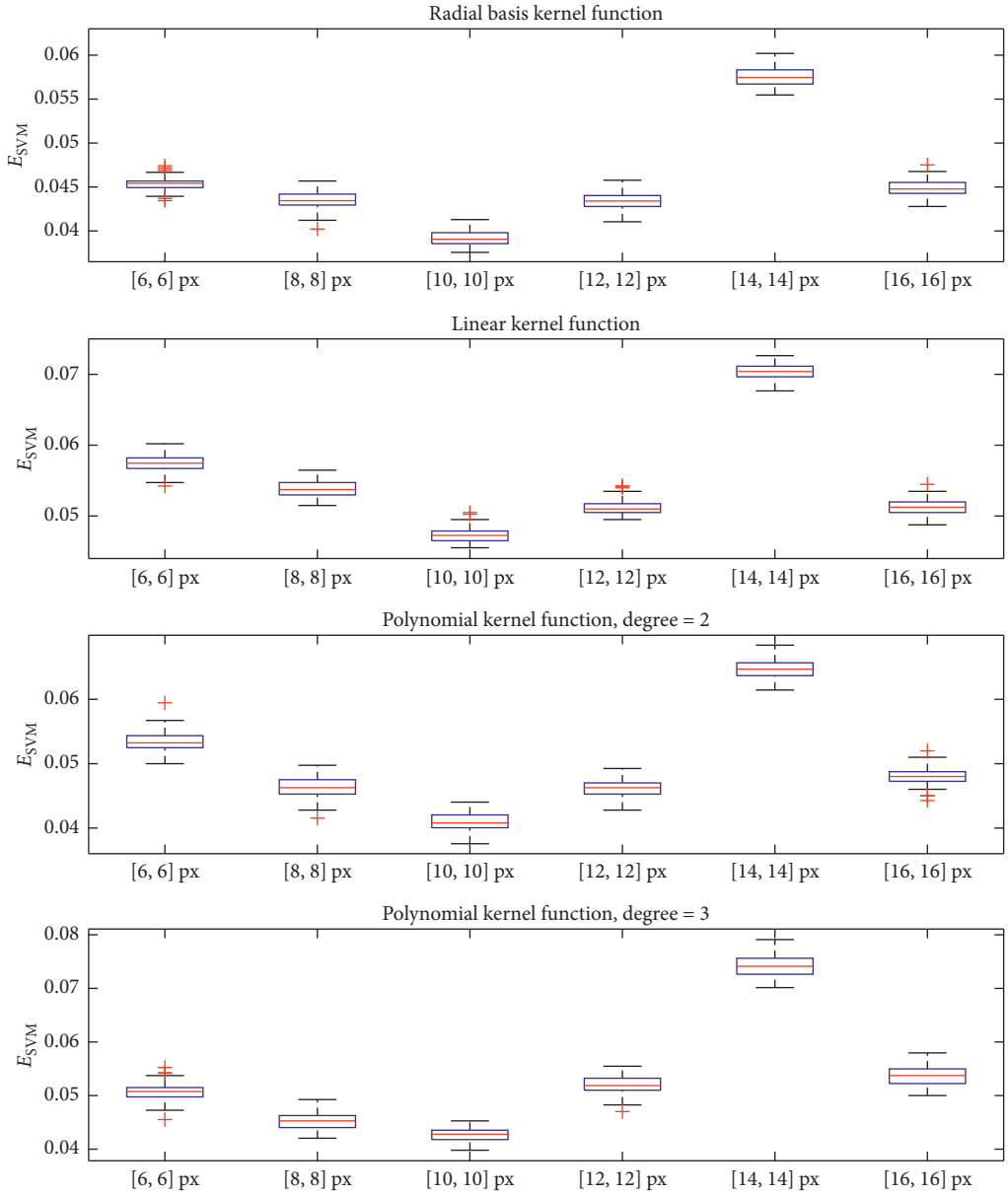


Figure 11: Boxplot representation of loss function values (y-axis) obtained by validation of the HOG-based passenger recognition system. Results are displayed in separate graphs with respect to used kernel function. In each graph, the data are grouped according to the sizes of cells (x-axis).

Table 3: Evaluation of classification performance of ConvNet-based (first section) and HOG-based (second section) passenger recognition systems using the measures (3)–(6).

Classifier	Accuracy	Precision	Recall	$F1$-score
Net1	0.949	**0.950**	0.948	0.949
Net2	0.953	0.947	0.961	0.954
LeNet-5	**0.956**	0.946	0.966	**0.956**
AlexNet	0.947	0.921	**0.977**	0.948
VGG_16	0.928	0.903	0.958	0.930
SVM with RBF kernel function, cell size [6, 6] px	0.949	**0.957**	0.941	0.949
SVM with linear kernel function, cell size [6, 6] px	0.939	0.947	0.931	0.939
SVM, polynomial degree = 2, cell size [6, 6] px	0.946	0.946	0.947	0.946
SVM, polynomial degree = 3, cell size [6, 6] px	0.953	0.953	0.947	0.950
SVM with RBF kernel function, cell size [8, 8] px	0.949	0.952	0.945	0.948
SVM with linear kernel function, cell size [8, 8] px	0.938	0.941	0.934	0.937

Table 3: Continued.

Classifier	Accuracy	Precision	Recall	*F*1-score
SVM, polynomial degree = 2, cell size [8, 8] px	0.947	0.951	0.943	0.947
SVM, polynomial degree = 3, cell size [8, 8] px	0.948	0.949	0.946	0.948
SVM with RBF kernel function, cell size [10, 10] px	0.956	0.964	0.947	0.956
SVM with linear kernel function, cell size [10, 10] px	0.943	0.946	0.939	0.943
SVM, polynomial degree = 2, cell size [10, 10] px	0.947	0.947	0.948	0.948
SVM, polynomial degree = 3, cell size [10, 10] px	**0.959**	**0.957**	**0.961**	**0.959**
SVM with RBF kernel function, cell size [12, 12] px	0.953	0.956	0.948	0.952
SVM with linear kernel function, cell size [12, 12] px	0.935	0.939	0.930	0.934
SVM, polynomial degree = 2, cell size [12, 12] px	0.950	0.955	0.945	0.950
SVM, polynomial degree = 3, cell size [12, 12] px	0.950	**0.957**	0.942	0.949
SVM with RBF kernel function, cell size [14, 14] px	0.929	0.925	0.913	0.919
SVM with linear kernel function, cell size [14, 14] px	0.919	0.925	0.913	0.919
SVM, polynomial degree = 2, cell size [14, 14] px	0.929	0.936	0.922	0.929
SVM, polynomial degree = 3, cell size [14, 14] px	0.921	0.920	0.923	0.922
SVM with RBF kernel function, cell size [16, 16] px	0.952	0.955	0.949	0.952
SVM with linear kernel function, cell size [16, 16] px	0.942	0.949	0.934	0.941
SVM, polynomial degree = 2, cell size [16, 16] px	0.948	0.951	0.944	0.948
SVM, polynomial degree = 3, cell size [16, 16] px	0.943	0.943	0.943	0.943

Note: best results are in bold.

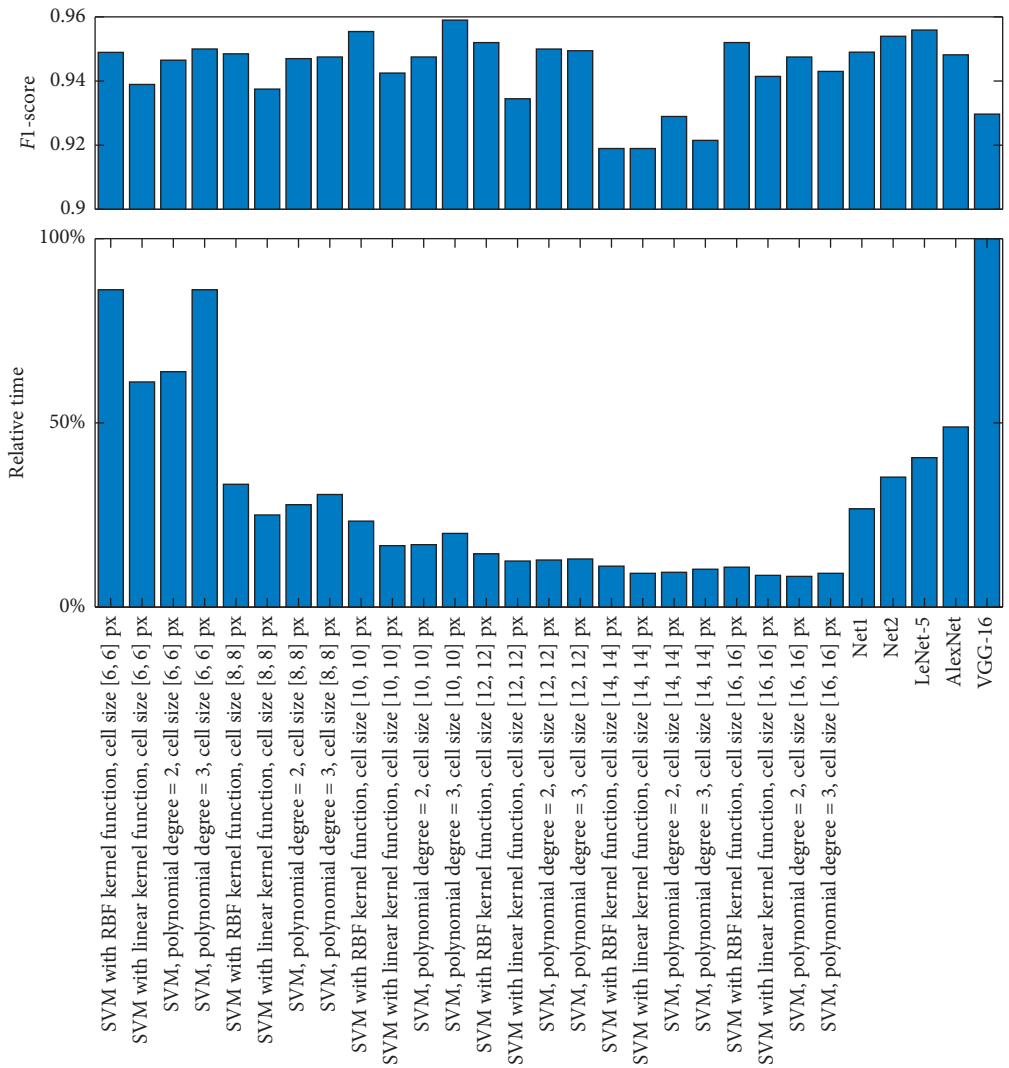


Figure 12: Relative computational times of the passenger recognition systems in comparison with *F*1-score of each system.

setting, the HOG-based system has significantly lower time complexity when compared to the ConvNet-based systems (Figure 12). Considering all these facts, we conclude that the HOG-based passenger recognition system, with polynomial kernel function of degree 3 and cells of size [10, 10] px, best fits the requirements for implementation into the low-cost automated real-time passenger counting system. This is in agreement with an earlier study of passenger recognition without the height differential [19].

The well-established ConvNets such as AlexNet and VGG-16 are expected to be a good basis of object recognition systems. As the validation results (Figure 10) show, they feature good learning ability, resulting in small loss function values. A similar ability can be observed for the AlexNet. From this perspective, the proposed networks Net1 and Net2 seem to be insufficient. However, their classification performance evaluated on the testing set (Table 3) is comparable with AlexNet- and LeNet-5-based systems (there is no clear winner among these four networks). Surprisingly, the VGG-16-based system has the worst performance in the category of the ConvNet-based systems. The most likely explanation of this phenomenon is a relatively high learning capacity of VGG-16 (compared to the other presented architects) that may cause overfitting on the head recognition task. Considering the high time complexity of VGG-16 (Figure 12), we conclude that, despite expectations, VGG-16 is not appropriate for the passenger recognition.

We also investigated possible benefits of the transfer learning by the training of ConvNet-based passenger recognition systems. We observe a lower variability in the cost function values for the networks trained using TL, when compared to the networks trained from scratch (Figure 10). Also, the medians of the cost function values are shifted towards smaller values for TL. We conclude that a model with a low cost function value can be more likely obtained using TL than by its training from scratch.

The size of cells has been reported to be the seminal parameter predetermining the performance of object recognition systems which are based on HOG features [27]. The experimental results presented in this article confirm this finding. An incorrect setting of the cell size results in inferior classification (compare results obtained for cells of size [10, 10] px and [14, 14] px in Figure 11 and Table 3). Also, the time complexity of the HOG-based system strongly depends on the setting of this parameter (compare, e.g., results for cells of size [6, 6] px and [10, 10] px in Figure 12).

5. Conclusions

Presently, deep ConvNets are usually considered as the first choice when developing an image recognition system. We established that image recognition systems with equally good classification performances can be developed using traditional computer vision methods. When appropriately designed and setup, such systems can beat ConvNets-based solutions in terms of time efficiency which is particularly important in real-world applications. This is also the case of the HOG-based passenger recognition system, where the utilization of HOG features in combination with the SVM classifier can result in time-efficient and accurate passenger recognition. In this context, we showed that passenger heads are sufficient for the precise while fast passenger recognition. We also showed that the HOG-based system is highly flexible, as it can be employed in both low-floor and high-floor means of transport. Its implementation into a passenger monitoring system is being currently developed, allowing us utilization of a basic processing unit. Cost savings on the unit is reflected in the final price of the person flow monitoring system and thus supports its mass use in means of transport all over the world.

Data Availability

Data used to support the findings of the study are available at https://www.researchgate.net/publication/342888989_Dataset_for_head_detector.

Conflicts of Interest

The authors declare that they do not have any commercial or associative interest that represents conflicts of interest in connection with the work submitted.

Acknowledgments

This work was supported from ERDF/ESF “Cooperation in Applied Research between the University of Pardubice and companies, in the Field of Positioning, Detection and Simulation Technology for Transport Systems (PosiTrans)” (no. CZ.02.1.01/0.0/0.0/17_049/0008394).

References

[1] A. Olivo, G. Maternini, and B. Barabino, “Empirical study on the accuracy and precision of automatic passenger counting in european bus services,” *Open Transportation Journal*, vol. 13, no. 1, pp. 250–260, 2020.

[2] M. Siebert and D. Ellenberger, “Validation of automatic passenger counting: introducing the T-test-induced equivalence test,” *Transportation*, 2019.

[3] J. W. Choi, X. Quan, and S. H. Cho, “Bi-directional passing people counting system based on IR-UWB radar sensors,” *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 512–522, 2018.

[4] Z. Chen, W. Yuan, M. Yang, C. Wang, and B. Wang, “SVM based people counting method in the corridor scene using a single-layer laser scanner,” in *Proceedings of the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, Rio de Janeiro, Brazil, November 2016.

[5] S. Akamatsu, N. Shimaji, and T. Tomizawa, “Development of a person counting system using a 3D laser scanner,” in *Proceedings of the 2014 IEEE International Conference on Robotics and Biomimetics (ROBIO 2014)*, Bali, Indonesia, December 2014.

[6] A. Ahmed and N. A. Siddiqui, “Design and implementation of infra-red based computer controlled monitoring system,” in *Proceedings of the 2005 Student Conference on Engineering Sciences and Technology*, Karachi, Pakistan, August 2005.

[7] T. Teixeira, G. Dublon, and A. Savvides, “A survey of human-sensing: methods for detecting presence, count, location,

track, and identity," Technical report, Yale University, New Haven, CT, USA, 2010.

[8] M. Mohaghegh and Z. Pang, "A four-component people identification and counting system using deep neural network," in *Proceedings of the 2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, Nadi, Fiji, December 2018.

[9] V. A. Sindagi and V. M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognition Letters*, vol. 107, 2018.

[10] S. Sun, N. Akhtar, H. Song, C. Zhang, J. Li, and A. Mian, "Benchmark data and method for real-time people counting in cluttered scenes using depth sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3599–3612, 2019.

[11] L. Del Pizzo, P. Foggia, A. Greco, G. Percannella, and M. Vento, "Counting people by RGB or depth overhead cameras," *Pattern Recognition Letters*, vol. 81, pp. 41–50, 2016.

[12] L. Sun, Y. Liu, S. Chen, B. Luo, Y. Li, and C. Liu, "Pig detection algorithm based on sliding windows and PCA convolution," *IEEE Access*, vol. 7, pp. 44229–44238, 2019.

[13] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[14] W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot multibox detector," in *Computer Vision—ECCV 2016*, pp. 21–37, Springer International Publishing, Cham, Switzerland, 2016.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[16] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017.

[17] X. Wu, "Design of person flow counting and monitoring system based on feature point extraction of optical flow," in *Proceedings of the 2014 5th International Conference on Intelligent Systems Design and Engineering Applications*, Changsha, China, June 2014.

[18] A. Tokta and A. K. Hocaoglu, "A fast people counting method based on optical flow," in *Proceedings of the 2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, Malatya, Turkey, September 2018.

[19] M. Belloc, S. A. Velastin, R. Fernandez, and M. Jara, "Detection of people boarding/alighting a metropolitan train using computer vision," in *Proceedings of the 9th International Conference on Pattern Recognition Systems (ICPRS 2018)*, Valparaiso, Chile, May 2018.

[20] T. Ma, Q. Ji, and N. Li, "Scene invariant crowd counting using multi-scales head detection in video surveillance," *IET Image Processing*, vol. 12, no. 12, pp. 2258–2263, 2018.

[21] M. B. Shami, S. Maqbool, H. Sajid, Y. Ayaz, and S.-C. S. Cheung, "People counting in dense crowd images using sparse head detections," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 9, pp. 2627–2636, 2019.

[22] S. I. Cho and S.-J. Kang, "Real-time people counting system for customer movement analysis," *IEEE Access*, vol. 6, pp. 55264–55272, 2018.

[23] J. Lee and B. Al, "Chapter 19—video surveillance," in *The Essential Guide to Video Processing*, pp. 619–651, Academic Press, Boston, MA, USA, 2009.

[24] P. Sharma and A. Singh, "Era of deep neural networks: a review," in *Proceedings of the 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Delhi, India, 2017.

[25] Y. Xu, X. Zhou, S. Chen, and F. Li, "Deep learning for multiple object tracking: a survey," *IET Computer Vision*, vol. 13, no. 4, pp. 355–368, 2019.

[26] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Washington, DC, USA, June 2005.

[27] P. Škrabánek and F. Majeník, "Detection of grapes in natural environment using HOG features in low resolution images," *Journal of Physics: Conference Series*, vol. 870, no. 1, Article ID 012004, 2017.

[28] D. J. Lope and A. Dolgun, "Measuring the inequality of accessible trams in Melbourne," *Journal of Transport Geography*, vol. 83, Article ID 102657, 2020.

[29] J. Preston, "Big buses in a small country: the prospects for bus services in Wales," *Research in Transportation Economics*, vol. 59, pp. 379–387, 2016.

[30] A. Kathuria, M. Parida, Ch. Ravi Sekhar, and A. Sharma, "A review of bus rapid transit implementation in India," *Cogent Engineering*, vol. 3, no. 1, 2016.

[31] P. Skrabanek and F. Majerik, "Evaluation of performance of grape berry detectors on real-life images," in *Proceedings of the 22nd International Conference on Soft Computing*, Brno, Czech Republic, 2016.

[32] P. Škrabánek and P. Doležel, "Robust grape detector based on SVMs and HOG features," *Computational Intelligence and Neuroscience*, vol. 2017, Article ID 3478602, 17 pages, 2017.

[33] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[34] Basler, "Basler ace," 2020, https://www.baslerweb.com/en/products/cameras/area-scan-cameras/ace/aca2500-60uc/.

[35] Computar, "Computar lenses," 2020, https://computar.com/product/705/M3514-MP.

[36] NVIDIA, "Nvidia jetson," 2020, https://developer.nvidia.com/embedded/jetson-tx2.

[37] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks," in *Proceedings of the 2017 International Conference on Communication and Signal Processing (ICCSP)*, Chennai, India, April 2017.

[38] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[39] L. Bottou, C. Cortes, J. S. Denker et al., "Comparison of classifier methods: a case study in handwritten digit recognition," in *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, Jerusalem, Israel, October 1994.

[40] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, https://arxiv.org/abs/1409.1556.

[43] P. D. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.

[44] ITU-R Recommendation BT 601, *Studio Encoding Parameters of Digital Television for Standard 4 : 3 and Wide Screen 16 : 9 Aspect Ratios*, ITU, Geneva, Switzerland, 2011.

[45] C. H. Lampert, "Kernel methods in computer vision," *Foundations and Trends® in Computer Graphics and Vision*, vol. 4, no. 3, pp. 193–285, 2008.

[46] Y. Sasaki and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, no. 4, pp. 427–437, 2007.

[47] J.-J. Zhang and P. Zhong, "Learning biased SVM with weighted within-class scatter for imbalanced classification," *Neural Processing Letters*, vol. 51, no. 1, pp. 797–817, 2020.

[48] P. Dolezel and D. Stursa, "Dataset for head detector," 2020, https://www.researchgate.net/publication/342888989_Dataset_for_head_detector.